Lab 09: Linear Regression Solutions

Team Name: Team member 1, Team member 2, Team member 3, Team member 4

2021-11-04

Load packages

```
library(tidymodels)
library(tidyverse)
```

Exercise 0

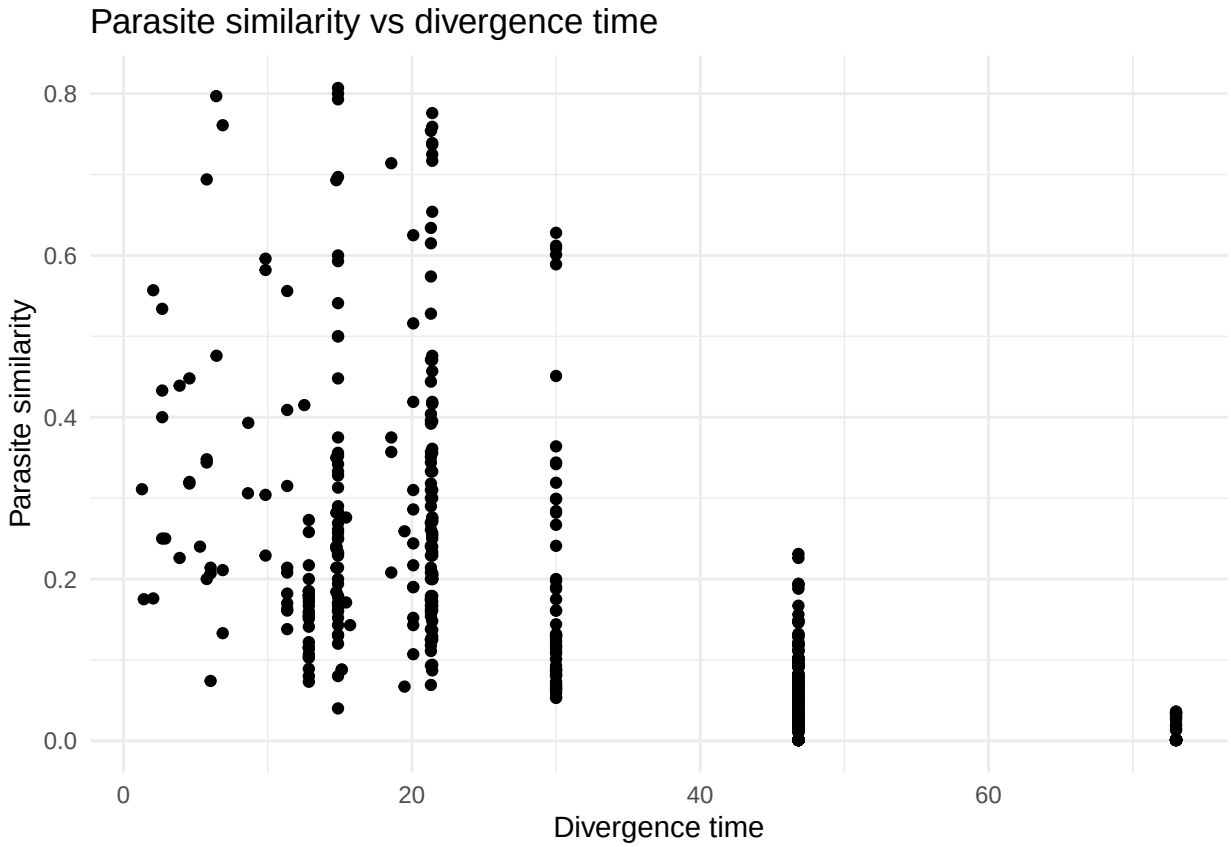
```
parasites = read_csv("data/parasites.csv")
parasites
```

```
## # A tibble: 595 x 7
##    species1        species2      divergence_time distance BMdiff precdiff parsim
##    <chr>           <chr>                   <dbl>    <dbl>  <dbl>    <dbl>  <dbl>
##  1 Alouatta_caraya Alouatta_gua~            4.57     932.   389.    57.5   0.318
##  2 Alouatta_caraya Alouatta_pal~            3.89    3456.  1000.    54.9   0.226
##  3 Alouatta_caraya Alouatta_sen~            2.89    1908.   821.     9.53  0.25
##  4 Alouatta_caraya Aotus_trivir~           21.3     1497.  4665.    27.6   0.154
##  5 Alouatta_caraya Ateles_geoff~           14.8     3785.  2005.    54.9   0.24
##  6 Alouatta_caraya Ateles_panis~           14.8     1422.  3120.    16.7   0.214
##  7 Alouatta_caraya Callithrix_j~           21.3     1739.  5287.     3.37  0.125
##  8 Alouatta_caraya Cebus_apella            21.3     1055.  2819.    11.4   0.29
##  9 Alouatta_caraya Cebus_capuci~           21.3     3109.  2571.    52.9   0.176
## 10 Alouatta_caraya Cercocebus_g~           46.8    10702.  1501.     2.84  0.001
## # ... with 585 more rows
```

Exercise 1

The outcome variable is `parsim`, parasite similarity.

```
parasites %>%
  ggplot(aes(x = divergence_time, y = parsim)) +
  geom_point() +
  labs(title = "Parasite similarity vs divergence time", x = "Divergence time", y = "Parasite similarity
  theme_minimal()
```

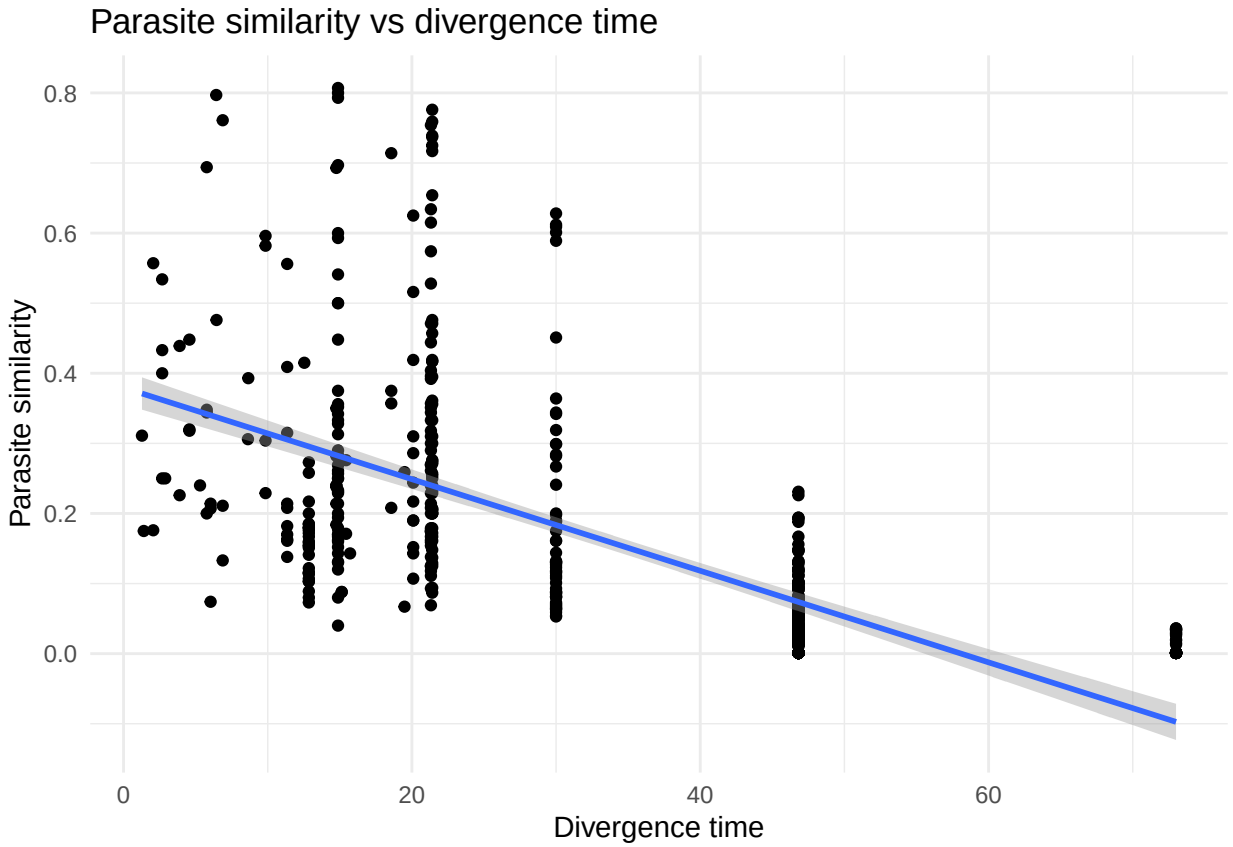


As divergence time increases, parasite similarity appears to decrease.

Exercise 2

```
parasites %>%
  ggplot(aes(x = divergence_time, y = parsim)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title = "Parasite similarity vs divergence time", x = "Divergence time", y = "Parasite similarity
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Model: $y = \beta_0 + x\beta_1$

y: parasite similarity

x: divergence time

β_0: intercept (i.e. if no divergence time)

β_1: effect of x_1, divergence time on parasite similarity

The regression line predicts negative parasite similarity for very large divergence times.

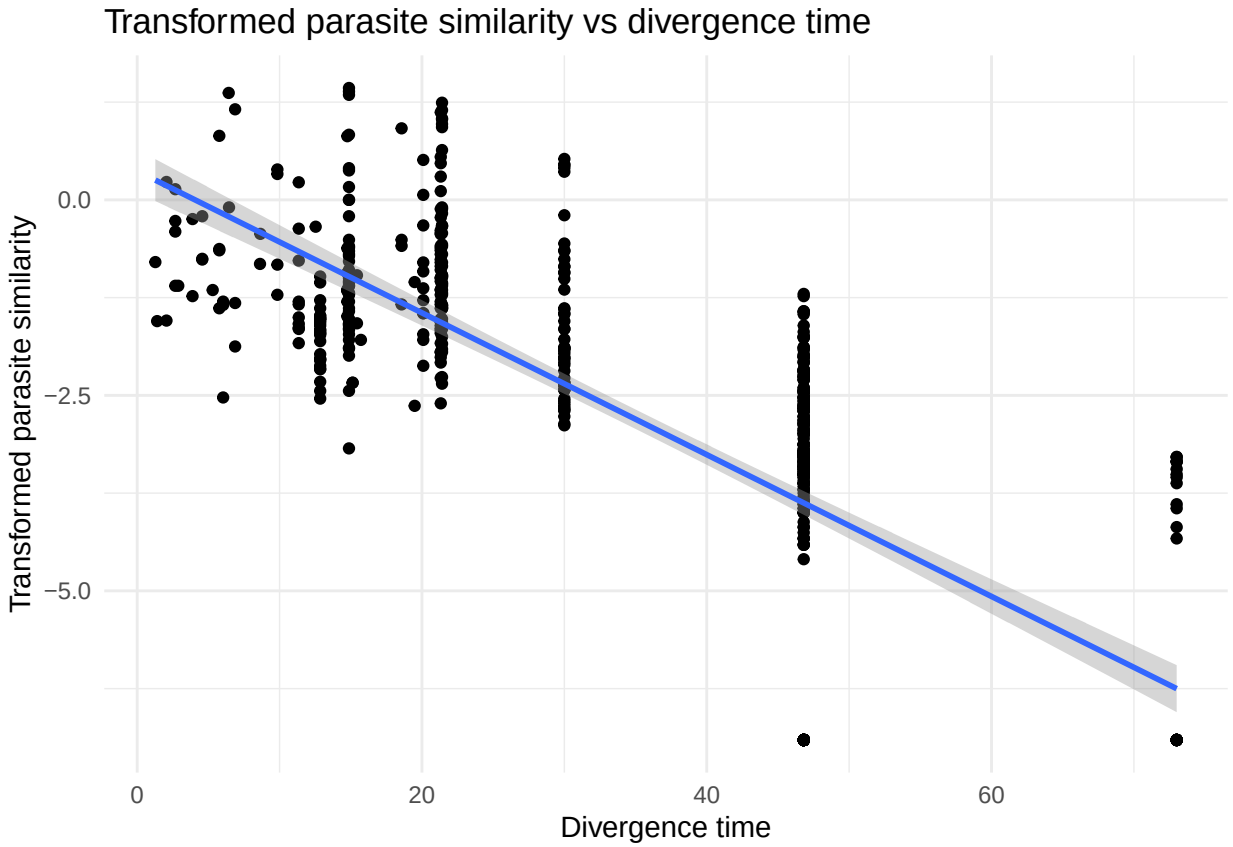
Also, (bonus): we would expect the intercept to be 1 if the species had spent 0 time diverging.

Exercise 3

```
parasites = parasites %>%
  mutate(transformed_parsim = log(parsim/(1-parsim)))
```

```
parasites %>%
  ggplot(aes(x = divergence_time, y = transformed_parsim)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title = "Transformed parasite similarity vs divergence time", x = "Divergence time", y = "Transf
  theme_minimal()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



It looks the same, but now the Parasite similarity values are not nonsensical and could be transformed back to meaningful numbers.

Exercise 4

```
dt_model = linear_reg() %>%
  set_engine("lm") %>%
  fit(transformed_parsim ~ divergence_time, data = parasites)

dist_model = linear_reg() %>%
  set_engine("lm") %>%
  fit(transformed_parsim ~ distance, data = parasites)

BM_model = linear_reg() %>%
  set_engine("lm") %>%
  fit(transformed_parsim ~ BMdiff, data = parasites)

prec_model = linear_reg() %>%
  set_engine("lm") %>%
  fit(transformed_parsim ~ precdiff, data = parasites)
```

Intercept represents estimated parasite similarity with 0 divergence time, geographic distance, body mass difference and precipitation difference respectively. In divergence time case the slopes show the parasite similarity if the two species were the same. (It's actually troublesome from a practical view that the intercept is not 1 in exercise 1)

The second estimate is the effect of each explanatory variable on the outcome parasite similarity. Specifically, it describes how much moving one unit in explanatory variable space affects moving one unit in outcome

space.

It is not useful to compare estimates between models because the variables are on different scales.

Exercise 5

```
glance(dt_model)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.513         0.512  1.54      623. 1.38e-94     1 -1102. 2209. 2223.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(dist_model)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.443         0.442  1.65      472. 1.95e-77     1 -1141. 2289. 2302.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(BM_model)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1   0.00327       0.00159  2.21      1.95   0.164     1 -1315. 2635. 2648.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(prec_model)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1   0.00861       0.00694  2.20      5.15  0.0236     1 -1313. 2632. 2645.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Divergence time is the strongest observed predictor of parasite similarity because it has the highest R^2.